Problems in Radiological Interpretation

J. YERUSHALMY, Ph.D., *Berkeley**

SUMMARY

Recent comparative x-ray studies demonstrate: (1) Photofluorography is a relatively efficient tool for tuberculosis case finding. (2) All photofluorograms should be interpreted independently by two competent interpreters. (3) The activity of a lesion cannot be determined from a single roentgenogram.

There is a variation in film interpretation between different observers, similar to that noted in many other fields which have been subjected to valid test. This variation is now the source of extensive investigations, out of which should come considerable progress in all branches of medicine.

THE recent intensive activities in mass radiography present to radiologists and chest specialists problems in diagnosis which are different in one essential element from those encountered in clinical work. In the latter, medical diagnosis is derived by a process of integration of a number of related observations and results of different kinds of tests obtained by repeated studies of the case. If a radiological examination is made in the course of such studies, the x-ray findings form only one link in a chain of pertinent evidence provided by physical examination and by other tests. In mass radiography, however, the physician is called upon to provide at least a tentative diagnosis on the basis of the appearance of the x-ray alone.

Since mass radiography is conducted among apparently healthy individuals, the persons examined are, on the whole, free from symptoms. In the majority of cases, the shadow on the x-ray film is the only basis for decision whether the patient is to be recalled for further study and how frequent and intensive the follow-up should be. In other words, mass radiography acquired certain functions of the primary diagnostician, which are different from those of a diagnostic aid to which clinical radiology is limited. It is doubtful whether any other branch of medicine is called upon to perform the functions of diagnosis with such limited observations.

In performing these diagnostic functions, the radiologist and chest specialist are further handicapped by the lack of experience on the probable prognosis of the types of lesions observed in mass radiography. These are more often of minimal extent and more fibrotic in character than those encountered in cases referred because of symptoms. Since mass radiography is a relatively new activity, not enough experience has been accumulated on which a comprehensive policy of treatment can be based.

In view of these limitations it is, perhaps, not surprising to find that diagnosis in mass radiography is not so definite as may be desired and that considerable variation can be demonstrated to exist in the interpretations of the same set of films by different readers and even in independent interpretations by the same reader. The existence of these variations poses a number of practical problems which can only be answered by intensive study of the many phases of diagnosis by means of x-ray. The main purpose of such investigations is to discover methods of improving and supplementing current practices of diagnosis in mass radiography in order to assist it in performing more efficiently the extraordinary task which has been placed upon it.

The problems of diagnosis by means of x-ray may conveniently be separated into two components. The first relates to the ability of the interpreter to detect "positive" cases—that is, his ability to separate photofluorograms into two groups, one which contains as nearly as possible all films with x-ray evidence of tuberculosis, and the other which contains all films which show no such evidence. The second problem relates to the ability of the interpreter to prognosticate from the appearance of the x-ray film shadow the probable severity and the course of the disease.

Theoretically, it may be argued that once the first task has been performed and the persons with x-ray evidence of tuberculosis have been selected, radiology has completed its task and the second problem becomes one for the clinician. In practice, however, it is found that the appearance of the x-ray film, and especially such changes in the appearance as may occur with the passage of time, forms the only available evidence of disease on a relatively large proportion of the cases and must therefore be relied upon almost exclusively for decisions as to the course of treatment.

Several attempts to study the two phases of this problem have been made in the last few years by a number of leading radiologists and chest specialists in the country. Recently, a group of radiologists in California have undertaken similar studies. Although no positive results have been attained to date, the problems, at least, have been more clearly defined. It may therefore be worthwhile to present a brief review of the methods and results of these investigations and such tentative conclusions that may be

*Professor of Biostatistics, School of Public Health, University of California, Berkeley.

From the Division of Biostatistics, School of Public Health, University of California and the Office of Field Studies, Tuberculosis Control Division, U. S. Public Health Service.

Presented before the section on Public Health at the 77th Annual Session of the California Medical Association, San Francisco, April 11-14, 1948.

derived from them which could be of assistance in mass photofluorography.

THE EFFECTIVENESS OF PHOTOFLUOROGRAPHY FOR TUBERCULOSIS CASE FINDING

The problems of variation in x-ray interpretation came to the fore in the course of an investigation on the relative effectiveness, for tuberculosis case finding, of various photofluorographic and roentgenographic methods. In this study, some 1,200 persons comprising the entire population of a Veterans' Administration institution were x-rayed consecutively on four different x-ray machines, providing for each person a 35 mm. photofluorogram of the chest, a 4 by 10 inch stereophotofluorogram, a roentgenogram on 14 by 17 inch paper negative and a conventional 14 by 17 inch celluloid film.

The four sets of films were interpreted independently by five experienced radiologists and chest specialists, members of a special board of roentgenology appointed by the Veterans' Administration. In addition, a second independent interpretation of the set of 14 by 17 inch celluloid films was obtained from each of the five members of the Board.

The original approach in that study was to compare the performance of the miniature and paper methods with that of the 14 by 17 inch celluloid films. In the course of the analysis, however, it was found that it was not possible to use the interpretations of the 14 by 17 inch celluloid films as a standard because of the great variation which exists, even on this method, for the different readers and between the independent interpretations of the same reader. This finding accentuates two problems. First, the lack of a standard makes it difficult to compare the effectiveness of the different photofluorographic methods. Second, it becomes important to investigate the implications to mass photofluorography of the inter-individual and intra-individual variations in film interpretation.

The first problem may be formulated as follows: Since the interpretations of the 14 by 17 inch celluloid films are not as unique and definite as previously supposed, it is not possible to determine who in the examined population is "positive" and who is "negative."* Such a determination is, however, essential before the different photofluorographic methods can be compared for their effectiveness in tuberculosis case finding. For, if it is not known who, in the examined population, should be selected, it is not possible to test the effectiveness of a given method in selecting "positive" cases. This problem has been solved by a method of comparison which was not dependent on the 14 by 17 inch celluloid films alone. The details of the method may be found in previous reports of this investigation.[1, 2] Briefly, the method utilized the evidence provided by all four methods, through the interpretations of all five readers for defining the persons examined as "positive" or "negative." By this method it was possible to compare objectively all the methods without a preconception of any one of them as a standard. In other words, each person in the study had 20 chances to be called "positive" since there were five readings on him for each of four films. A person, therefore, was considered to have x-ray evidence of tuberculosis if at least 11 of the 20 readings were positive for tuberculosis. The different methods were then compared according to the number of positive readings which each of them contributed on these "positive" cases. It might be expected that the more efficient method would contribute more positive readings on these cases than one which is less efficient. It was found that the number of positive readings was nearly the same for each of the techniques.

On the basis of this finding it was concluded that, strictly from the standpoint of their effectiveness in finding cases of tuberculosis, none of the methods, not even the 14 by 17 inch celluloid, is superior to any of the other methods.

It is probable that the second problem, that of variations in film interpretation, was a contributing factor to this finding. In other words, the result may also be stated as follows: The subjective error in film interpretation is so large, that compared with it, such differences in superiority as may exist between the different methods is small enough to be negligible.

The subjective error as measured by variation in film interpretation was found in that study to be of the order of magnitude of approximately 25 per cent. In other words, nearly a quarter of the films which were called positive by one reader were called negative by another. The degree of inconsistency of each reader with himself was found to be of approximately the same order of magnitude.

SUBJECTIVE ERRORS IN FILM INTERPRETATION

The unexpectedly large subjective error in film interpretation observed in the Veterans' Administration study was disturbing to a number of radiologists. Doctor W. Edward Chamberlain, who participated in the Veterans' Administration study, thought that the subject was deserving of further intensive study and at his initiative a board of radiology was appointed by the U. S. Public Health Service to conduct such an investigation. The members of the board are Dr. Chamberlain of Philadelphia, Dr. Leo Rigler of Minneapolis and Dr. Robert Newell of San Francisco. The board, at the time this presentation was prepared, had been at this problem for nearly two years. It is not my intention here to "steal their thunder" and present to you a premature account of their findings. I know that a comprehensive report of the results of that study will be published in full. I may, however, indicate in very general terms the trend of their findings.

In the first place, in their initial trials, the magnitude of the variations in calling a film "positive" or "negative" was found to confirm the results of the

*The terms "positive" and "negative" as used in this paper do not refer to proven cases of tuberculosis, but to persons whose x-rays show or do not show shadows suggestive of tuberculosis.

Veterans' Administration study. In fact, on a set of 100 films, on which are available the results of 20 independent interpretations of ten radiologists and chest specialists throughout the country, similar findings were observed. As a curiosity it may be worth mentioning that at the last count of the readings on these 100 films which were selected to contain approximately 33 "positive" films, only fourteen were called "negative" by all the readers, and for fully 93 films there is at least one positive reading by one or more of the ten readers.

The board concentrated its efforts on attempts to devise methods of film interpretation and on the development of nomenclature which would lead to greater consistency. All the accepted terms used in radiology of the chest were carefully tested and found to be unreliable to a greater or lesser degree. These run the entire gamut from the National Tuberculosis Association classification of minimal, moderately advanced and far advanced, through attempts at simple description of the texture of a lesion. It was found that it is apparently not possible to determine the activity of a lesion from the appearance of the x-ray shadow, since too many lesions which are called active by one reader will be called inactive by another. Moreover, the same reader is likely to call a lesion inactive when he has previously called it active. Similar limitations apply when a lesion is described as "soft," "hard," "exudative," "productive" or "fibrotic." The same lesion may appear as "band-like" at one reading and as "fan-like," "linear" or "round" at another. It may be called "poorly defined" once and "well defined" another time. Similarly, no reasonable agreement can be obtained either among different readers or for a single reader with himself in describing a shadow as "homogeneous," "honeycombed" or "spotted."

In general, it may be stated, that while it may be possible by special effort and training to succeed in reducing somewhat the variation in calling a film "positive" or "negative," no appreciable progress has yet been made in getting a group of radiologists to use common language in describing what is seen on the x-ray.

It may not be out of place to say a few words here on the problem of serial films. The general impression is that all these problems arise from the fact that so much is asked from the interpretation of a single film, but, that if a set of films taken at different periods were available, no such problems would arise. This is not entirely true. In connection with these activities on film interpretation, the U. S. Public Health Service established long-range follow-up studies on persons discovered through mass radiography. These studies are being run concurrently with the radiological investigations because it is realized that the acid test of such schemes that will be developed in film interpretation lies in their correlation with prognosis. It is much too soon for any pertinent conclusions to be drawn from these studies but on the question of serial films considerable information is already available.

As soon as a set of two films on the same individuals became available and was sent to different readers for interpretation as to stability, progression, or regression, it became apparent that there is as much variation in the interpretation of a set of serial films as there is in that of a single film. Since then considerable experience has been accumulated on interpretations of sets of five or six films taken on the same individuals at three monthly intervals. It is revealing to observe several competent radiologists and chest specialists review the same set and arrive at different conclusions as to the stability of the lesion over a period of 15 to 18 months.

The difficulty in this case appears to result from technical differences in positioning, stage of respiration, film development and the like. It becomes a question of how much of the difference in the x-ray shadows the radiologist is ascribing to these technical factors and how much he thinks is due to changes in the lesion. The net result is that with the exception of very obvious progressions or regressions, it is difficult to determine from a series of films whether the lesion is active or inactive.

DISCUSSION

In studying the implications to photofluorography of the inter-individual and intra-individual variations in x-ray interpretation, it may be well to differentiate between ultimate and immediate objectives. Final solution to the basic problems of variation in x-ray diagnosis will evolve only as a result of fundamental research in the underlying causes for the phenomenon of variation. Many different branches of science: physics, psychology, medical optics, semantics, and many of the medical sciences, will, no doubt, supply important contributions to the final solution. In addition, more knowledge on the ultimate course of the disease and its correlation with the x-ray appearance will need to be accumulated and integrated with the purely radiological investigations. These studies, obviously, are long-range and require intensive effort.

The immediate objectives are to utilize such knowledge as is available in attempts to improve, as far as possible, present practices in radiology. It is not necessary to wait for the final solution for the development of methods of interpretation and tentative methods of classification which will reduce the variation in x-ray diagnosis. In fact, considerable improvement in the practice of photofluorography can be attained by the application of certain conclusions which may be drawn from the results of these studies to date.

In the first place, it is important not to lose sight of the main result of the Veterans' Administration study, which is, that for case-finding purposes it is just as safe to operate with the miniature methods as it is with the expensive and cumbersome 14 by 17 inch celluloid method. In view of the rather extensive activities in the field of photofluorography, this may sound like a superfluous assertion. Nevertheless, it is reassuring to know that the methods em-

ployed in mass case-finding programs are as efficient as the best available.

The second practical conclusion which may be drawn from a knowledge of the magnitude of variations in film interpretation is that a simple method is available to safeguard against much of the harm which may be caused by these variations. This can be attained by the simple procedure of having at least two independent interpretations of each set of films. If the probability of overlooking a positive film is approximately 1 in 4, then the probability of missing a positive film in two independent readings is 1 in 16, and that in three independent readings is only 1 in 64. It therefore becomes possible to obtain almost any degree of precision by the simple expediency of multiple reading. How many such readings are desirable is in the long run a matter of individual taste. To some persons the oversight of even one positive case is a matter of grave concern. Most of us, however, must take into consideration the practical problems involved in multiple readings. When approached from this practical point of view it can be demonstrated that double reading is actually an economical procedure. This may perhaps be illustrated by the following example: Suppose 10,000 persons were examined. If the prevalence of tuberculosis in this group is one per cent, there are 100 persons in the group who should have been selected. A single reading of the 10,000 films will uncover approximately 75 of them. By means of a second independent interpretation of these films, 18 or 19 of the remaining 25 will be discovered. In order to find 18 or 19 cases by other means, another group of some 2,000 individuals must be examined. It may be easily shown that the expense involved in the latter process is greater than that of a second reading of the original 10,000 films.

It is realized that it is relatively easier in these times to find technicians to operate mass survey projects than it is to secure the services of competent physicians to interpret the films. Nevertheless, it is false economy to conclude that the agency cannot afford the "luxury" of double reading. It is by no means a luxury. It is good dollars and cents economy.

The third conclusion which may be drawn concerns the type and intensity of follow-up indicated for cases discovered in mass photofluorography projects. Since it is impossible to determine from the appearance of the x-ray shadow the activity and probable prognosis of the lesion, it becomes important to keep under frequent and continuous observation the majority of cases discovered in these projects. It happens too often that a patient that one specialist would dismiss as requiring no further study would have been kept under close observation, and even hospitalized, if the opinion of another, equally competent, specialist would prevail. In fact it is likely that the first man would have called the same case active on another reading of the same x-ray. It follows, therefore, that at the present stage of knowledge, a conservative policy of keeping a large proportion of these cases under observation is desirable.

In addition greater efforts must be exerted to supplement the x-ray findings with other tests, particularly with bacteriological examinations. Since many of the patients do not raise sputum, special efforts must be exerted in securing samples of gastric lavage in as large a proportion of the cases as is possible. This may not be too easy, but judging from the experience in Denmark, the efforts will be greatly rewarded. For example, last year 136,000 persons 15-35 years of age in Copenhagen were x-rayed. As a result of these examinations, 344 persons were found with positive sputum or gastric washings. The striking fact is that in fully 72 per cent of these 344 persons the tubercle bacilli were isolated from samples of gastric lavage. In other words, in this country, where stomach washings are not very commonly used, only about one-fourth to one-third of the 344 patients would have been definitely diagnosed as having tuberculosis. In order to obtain full benefit of the case-finding programs, at least as much effort should be devoted to the follow-up of the cases as was exerted in discovering them.

In conclusion, it may be worth emphasizing that the demonstration of the variation in film interpretation need not be considered as a negative result. A knowledge of the magnitude of such variations may be used in a very positive way in determining many practical problems of policy in mass radiography. However, intensive and concentrated study is indicated for a real attempt at determining the fundamental problems of the underlying causes for these variations as well as of many other phases of the general problem of diagnosis by means of x-ray. A number of competent persons and groups throughout the country are engaged in such research and it may be hoped that through their efforts great progress will ultimately be achieved.

REFERENCES

1. Birkelo, C. C., Chamberlain, W. E., Phelps, P. S., Schools, P. E., Zacks, D., and Yerushalmy, J.: Tuberculosis case finding, J.A.M.A., 133:359-365, 1947.

2. Yerushalmy, J.: Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques, Public Health Report, 62:40:1432-1449, 1947.

Discussion by L. H. GARLAND, M.D.

The problem of mass radiography has as many facets as a well-cut diamond. One of these, the problem of reliable roentgen interpretation, has been ably outlined by Dr. Yerushalmy. In order that some of Dr. Yerushalmy's figures may be clearly understood, I think it is important to note that he uses the strict statistical approach. For example, if one of his "testing" radiologists or chest physicians is submitted 100 chest films for study, and these contain two "positives," the reader will have a score of 50 per cent inaccuracy if he misses *one* of those "positives," even though he read all the 98 "negatives" correctly. This point is noted in discussion and not defense, since the roentgen method requires not defense but clarification and intelligent usage. Errors resulting from the "personal equation" exist in all fields of knowledge; it is only in those in which the evidence is easily measurable

that its magnitude can be readily demonstrated. Berkson and associates have shown the astonishing errors in estimation of simple blood-cell counts; Derryberry has exposed the incredible discrepancies in interpretation of degrees of malnutrition. The roentgenogram is only as good as the person interpreting it, and therefore subject to the not inconsiderable limitations of human perception.

There are some specific points in Dr. Yerushalmy's paper upon which I would like to comment. He stresses wisely that mass radiography is a problem in diagnosis; the physician is called upon to provide at least a tentative diagnosis on the appearance of the roentgenogram. This is noted in order to correct the specious arguments that mass surveys are not a medical or diagnostic procedure.

There is a tendency on the part of many persons, even scientists like Dr. Yerushalmy, to use the expression "x-ray evidence of *tuberculosis*." What he really means is "x-ray evidence of *disease*." The bacterial origin of the shadow cannot be determined from a single or even from serial roentgenograms. This point he emphasized himself when he referred to the "great need for accurate and frequent sputum studies." Indeed, many years ago, when mass x-ray surveys were first being projected, I had the temerity to suggest that mass sputum surveys would be a more worthwhile public health measure, since it is the contagious case rather than the person with a silent shadow who is the true public health menace.

The fact that the activity of a lesion could not be determined from the appearance of the x-ray shadow alone has been known for over 40 years. We recollect being told this in 1925. Nevertheless, in everyday clinical practice we do believe that the activity of many lesions can be more accurately estimated by serial roentgen studies than by any other objective method.

Double reading of films, that is to say, separate, independent readings by two radiologists, greatly diminishes the possibility of overlooking a positive roentgenogram. This use of double reading in minifilm surveys is wisely stressed. The readers ought to use a consistent reading technique, and their remuneration for the service should be such that there would need be no tendency to give inadequate study to apparently negative films.